iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75910

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

A Hybrid MLSELM and Threat Intelligence Integration System for Real-Time Zero-Day Phishing Detection

Dr. Vairam T[1], Lakshmi Priya S[2]
[1]*PSG College of Technology, Coimbatore, India*
[2]*Master of Engineering, Biometric and Cybersecurity, PSG College of Technology*

Abstract: ***Phishing remains a significant cybersecurity issue with attackers exploiting social engineering to pressurize users into disclosing personally identifiable information. This paper provides a refined Chrome Extension which uses machine learning as well as real-time threat intelligence (TI) in its endeavors to ascertain phishing. The system employs the Multi-Layer Stacked Ensemble Learning Model (MLSELM) with TF-IDF-based features and gathers real-time data through threat intelligence APIs, including PhishTank and OpenPhish. The Flask server creates an interface that lets you make real-time alerts for multiple zero-day phishing URLs. This lets you use the ML model and combine the API outputs. This hybrid model makes it easier to respond to threats, lowers the risk of false negatives, and gives users better protection. Keyword- Phishing, Machine Learning, Threat Intelligence, Chrome Extension, MLSELM, Real-time Detection.***

I. INTRODUCTION

To different extents attackers change domain names, URLs and even the content of the website just to surpass the traditional security measures. The continuously improving ways of phishing make the traditional rule-based or blacklist detection systems absolutely impotent to the new and unseen phishing attempts. So, the adaptive, data-oriented defense mechanisms are going to be increasingly essential for real-time user protection, at least as the phishing tactics develop.

Machine learning (ML) through URL-based and content-based feature analysis, has become one of the most effective and powerful techniques for detecting phishing. The Random Forest and XGBoost, which are ensemble models, have reached outstanding heights in phishing URL classification. The Multi-Layer Stacked Ensemble Learning Model (MLSELM) applies TF-IDF-based feature extraction to fuse multiple classifiers and thus, bolsters the predictions. This model by the analysis of the words and the structure of URLs is able to tell apart the legitimate websites from casting ones. But, it is a challenge for ML models to spot the phishing URLs that have just come into existence and are not already included in training data.

To overcome this challenge, this project integrates real-time Threat Intelligence (TI) feeds into the existing MLSELM framework. Platforms such as PhishTank and OpenPhish continuously update verified phishing URLs reported by global users. The proposed hybrid model fuses ML predictions with TI responses through a Flask-based fusion logic layer. By lowering false negatives and increasing the system's flexibility against new threats, This integration helps detect new phishing domains in real time by reducing false negatives and enhancing the system's adaptability to new threats. The fusion ensures that any verified phishing URL is quickly flagged, even before the model is retrained.

The new system went live as a Chrome Extension and evaluates websites as its users browse the web and sends them real-time alerts. This extension with a backing Flask server carries out identical TI based and ML-based testing in parallel. Every URL is ultimately detected as either Safe, Suspicious, or Phishing depending on the overall evidence amount of each URL. The hybrid structure will offer reliable and scalable real-time cybersecurity to everyday users because it ensures great levels of accuracy, fast response, and proactive protection against evolving phishing attacks.

II. RELATED WORK

The article by Kalabarige et al. in 2022 provides a Multilayer Stacked Ensemble Learning Model to detect phishing, establishing an accuracy of between 96.79 and 98.90. In this case, the ensemble methods have been applied in a way that they surpass the conventional classifier. It appears that there is a growing danger of phishing, which the model architecture will discuss, and the performance will be thoroughly tested. This implies possession of good methods of detection and suggestions towards future works is the use of the better feature selection and optimization.

In [2], Aminu et al. have suggested an extensive structure named Enhancing Cyber Threat Detection via Real-time Threat Intelligence and Adaptive Defense Mechanisms. The article is concerned with the combination of real-time threat intelligence (TI) with adaptive machine learning models to reinforce cyber defenses. Their strategy allows uninterrupted tracking and dynamic reaction on the changing threats based on intelligence-driven data correlation. The model demonstrated high detection rates and fast response of zero-day attacks showed that TI is valuable in the real-life security systems. The authors noted that implementation of live intelligence feeds is one of the most important improvements in the speed and flexibility of detection. Further development and work is being done to automate the automation layer and work with the model on a scale of large enterprise settings.

Kalabarige et al. (2023) developed a hybrid boosting-based parametric selection and multi-layered stacked ensemble paradigm for phishing detection that reaches accuracy levels close to 98.95%. The benchmark models are much better than previous ones, which were purely through a hybrid selection method integrated into multi-layer stacked ensemble learning. This paper addresses the phishing threat extent, covering architectural or performance features of the model, as well as the need for selection of important features for optimization of detection. The study proposes a strong empirical methodology for demonstrating a strong phishing detection system and indicates future grounds for feature selection and model optimization.

Bell and Komisarczuk (2020) [4] conducted an analytical study comparing multiple open-source phishing blacklists such as Google Safe Browsing, PhishTank, and OpenPhish. Their work highlighted that although blacklist-based systems provide real-time updates, they suffer from limited coverage and higher false negatives when detecting newly created phishing sites. The paper has highlighted that the integration of dynamism threat feeds with other data-driven models would make resilience more effective. They came up with the conclusion that incorporating real-time intelligence sources would greatly enhance early mechanisms of phishing detection.

Patel and Bhat (2024) [5] suggested a hybrid phishing Franciscation model that includes machine learning classifiers with Threat Intelligence (TI) feeds of PhishTank. Their model is reaching the level of 98.7 overall accuracy and it can be stated that this proves that the system is less reliant on the fixed training data and TI-enhanced datasets. The article has introduced a hybrid-based detection model, in which TI-validated phishing sites were used as an authentication layer on top of the ML results. The authors hypothesized that real-time feed integration is given a high preference, as it enhances flexibility towards zero-day phishing attacks.

Ensemble machine learning with real-time threat intelligence feeds to detect and block in real time was proposed by Pradhan (2023) [6] as a Zero-Day URL Defense Model. The system used URL reputation scores based upon various APIs and included them as a weighted ensemble decision system. This mixed architecture obtained more than 98% of accuracy during the test on experiments and demonstrated a better memory of hitherto unseen phishing web sites. It was discovered in the research that this kind of integration would improve situational awareness and responsiveness of threats in a timely manner, minimizing false negatives by a considerable margin.

Adebayo et al. (2024) [7] developed a multi-source threat intelligence design that assembles and classifies the phishing threat by incorporating the deep neural networks (DNN). Their method obtained threat information on OpenPhish, PhishSTAT, and Abuse.ch and used them as training examples to the ML model. The system had a detection rate of 99.1 and better zero-day coverage rates by updating the database through dynamically verified phishing URLs. The authors have focused on the fact that such a hybrid form of integration allows a self-adaptive cybersecurity system that can learn by using continuous threat feeds.

Sahingoz et al., 2024 [8] have come up with a highly advanced phishing detection system they named DEPHIDES which employs the concepts of ANN, CNN, RNN, BRNN, and attention networks under the umbrella of deep learning. The study probes URL-based rapid classification through an extremely large dataset of 5 million labeled URLs. CNN has been found to achieve the highest accuracy of 98.74% among the rest of the models presented. Apart from indicating the looming phishing threat, the paper delineates quite profoundly the features of model architectures and performance metrics. Advanced deep learning methods have been said to be the prerequisites to improve detection. Further studies are said to be expected toward bettering feature selection and models' optimizations to get higher accuracy in outputs.

Almousa et al. (2023) [9] designed a model that detects social semantic attacks through the analysis of URLs that are aware of characters in their construction. Its phishing threat identification accuracy reached 96.5%, of which improved detection exploits were observed when considered at the level of individual characters through obfuscated URLs. Such experiments endorsed the skill of the model towards changing attacks. This work effectively derived meaning for semantic analysis about a cybersecurity issue. Future works are supposed to annotation the context that surrounds that information to enhance the detection of threats.

Zhou et al. (2023) [10] proposed an intelligent phishing detection framework that combines machine learning algorithms with Threat Intelligence correlation analysis. The system extracts URL-based features and cross-verifies them with live threat repositories using API calls for real-time validation. Their hybrid model achieved 99.03% accuracy and demonstrated improved resistance against adversarially generated phishing URLs.

The researchers highlighted that integrating TI feeds enhances model adaptability and significantly reduces false negatives. They concluded that combining ensemble ML with continuous intelligence updates offers a scalable defense strategy for real-world phishing prevention.

III. MODEL

A. *Existing System*

Phishing detection systems have mostly used machine learning (ML) and deep learning (DL) models to classify websites as either phishing or legitimate based on their URLs. They assess websites and identify malicious activity based on feature extraction, lexical analysis, and content analysis. They have commonly employed algorithms such as Random Forest, SVM, and XGBoost in order to attain high accuracy in static datasets. These methods only capture features of previous patterns in an observed dataset and are therefore based on prior data detection. These methods can simply miss zero-day phishing URLs not previously observed and therefore do not represent data generated by existing algorithms. One additional detection limitation in current systems is the acceptable use of black lists or rules, both of which become easily outdated. Ensemble models, or hybrid, ML models might improve accuracy (precision) but ultimately are not equipped with the ability to detect changing phishing threats dynamically in real-time. There is a need to integrate live Threat Intelligence (TI) sources to identify phishing threat actively.

B. *Proposed System*

The proposed work introduces a hybrid phishing detection framework that integrates machine learning model with threat intelligence to enhance resilience against sophisticated and evolving attacks. The system combines live threat feeds from phishtank and open phish and with an ensemble learning engine comprising XGboost and random forest enabling real time identification of malicious URL. By continuously validating prediction against active threat source the framework adapts to zero-day pattern and minimizes false positives. The hybrid design offers a proactive and reliable cybersecurity mechanism capable of delivering improved detection accuracy and timely protection against emerging phishing threats.

C. *Datasets*

The phishing detection system uses a hybrid dataset that was collected in the UCI Machine Learning Repository and Kaggle to ensure balance and diversification of training data. All the datasets contain URLs, which are labeled as legitimate or phishing. The data has lexical and structural features that facilitate the detection of phishing. Phishing examples in Live Phishing of PhishTank and OpenPhish were incorporated in order to enhance flexibility by simulating real-time zero-day conditions.

D. *Data Collection from Open Repositories*

The information was collected on the UCI ML Repository and Kaggle that provide real-life phishing datasets collected by browser crawler and user reporting. These datasets comprise thousands of labeled URLs that can be evaluated and trained in an efficient way. The ability to distinguish a large variety of phishing schemes is added to the model through the addition of different URL forms. Moreover, live data on OpenPhish and PhishTank enhances the effectiveness of detection of unknown phishing attacks.

E. *TF-IDF for Feature Extraction*

The system uses the Term Frequency-Inverse Document Frequency (TF-IDF) to extract important URL-based lexical features. The TF-IDF compares the unique patterns and that hardly appears in genuine URLs yet is widespread in the phishing URLs. These features are examined in the MLSELM ensemble to enhance the presence of the classification accuracy. This is supplemented by Threat Intelligence module, which provides effective and dynamic detection based on checking suspicious patterns using real-time phishing notices.

F. *Training the model*

1) *Multilayer Stacked Ensemble Model:* The system is used to identify phishing URLs by means of Multilayer Stacked Ensemble Learning Model (MLSELM) that would process both Random Forest and XGBoost. An amalgamation of the two classifiers allows one to make accuracy predictions more. TF-IDF features are used as inputs and the output is combined by means of a voting process. Dependability and zero-day adaptability are high since the cross-checking of the predictions against verified feeds provided by PhishTank and OpenPhish ensures this.

2) *XGboost Model:* To enhance the level of accuracy, XGboost, which is a gradient boosting algorithm, constructs sequential decision trees in a scalable and fast fashion. It is ideal in detecting phishing because it can handle large and unbalanced datasets. One of the primary models of MLSELM is XGboost which is individually tested in this system. Its findings are checked with data of Threat Intelligence and increase the legitimacy of the model and precision of its detection.
3) *Random Forest:* A Random Forest is a type of decision tree used to create results that are both reliable and accurate using various decision trees. It is also very precise with no overfitting and it works with noisy data. It provides a reliable performance through an additional classifier within the MLSELM. Random Forest enhances resilience and resistance to phishing attacks and new and hidden threats when used in conjunction with the Threat Intelligence verification.

G. Components

To implement real-time phishing detection based on a hybrid ML and threat intelligence solution, the proposed system architecture is a collection of a flask-based backend and a Chrome extension. The extension is easy to use and has a URL input interface with a backend, which can process data fast, with a restful API. With TF-IDF-based URL features, the backend employs ensemble MLSELM model which integrates XGboost and random forest. All URLs are checked with the help of phishtank and openphish live threat intelligent checks and ML. The two outputs are merged with a layer of fusion logic, which will give color-coded feedback, which is red (phishing), and green (safe), which will give real-time, precise, and adaptive protection against zero-day threats.

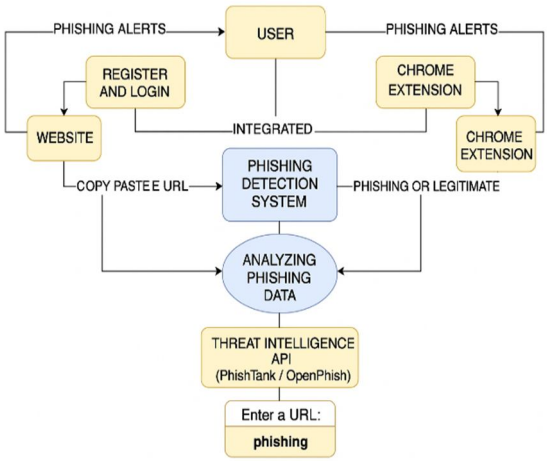


Figure 1 – Components

H. Architecture

1) Front end based

Users can either type in the URLs themselves or they have them automatically checked against phishing attacks since the front end will be designed as a Chrome Extension. It provides a simple and user-friendly interface and has an analysis API to the backend. Immediately the user can view the output of the processing of the URL with the Threat Intelligence Model with XGBoost and Random Forest: Safe or Phishing. Under browsing, this setup will ensure that it detects and provides real time feedback.

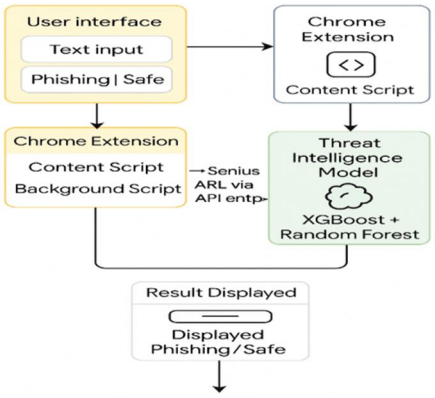


Figure 2 – Front-end based architecture diagram

2) Backend-based Architecture diagram

To detect phishing with high accuracy, the proposed system is a combination of Threat Intelligence (TI) and a hybrid ensemble learning approach. It is an effective URL pattern analyzer consisting of the Random Forest classifier and XGBoost classifier, which are combined in the MLSELM framework. TI module employs trusted sites, such as PhishTank and OpenPhish, to check suspicious URLs. This combination reduces erroneous detections, and puts on a general increase in reliability and a higher ability to adapt to the zero-day attacks.

This process starts by a user putting in a URL through the Chrome Extension interface. The content script captures the URL and post to the using a RESTful API to Flask backend. To be analyzed by the backend TF-IDF feature extraction process, one transforms the URL into numerical values. XGboost and Random Forest classes are used to handle such features in the MLSELM framework.

In a fusion layer, majority voting method is utilized to discover the fusion of the two models to get the final prediction. A concurrent Threat Intelligence Testing is also performed by the system on live feeds on PhishTank and OpenPhish. When it is detected that a URL is malicious, it is shown to be Phishing (Threat Verified). The results are then automatically shown as a Chrome Extension with green color-coded feedback indicating that the site is safe and a red color-code of a phishing site. This hybrid workflow provides real-time fast and reliable phishing detection that is flexible.

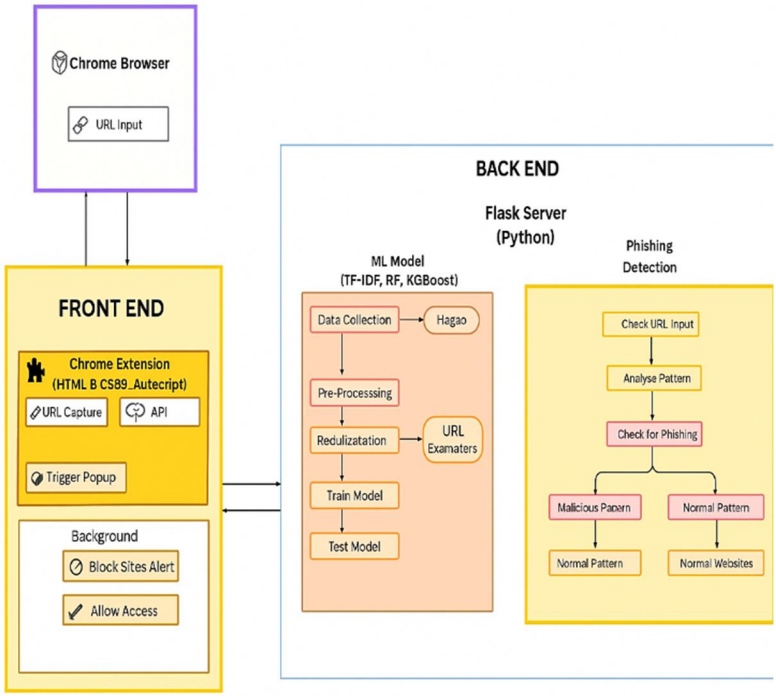


Figure 3 – Front-end based architecture diagram

IV. RESULTS ANALYSIS

The project analyzes three fundamental machine learning models MLSELM, XGboost and Random Forest (RF) in the activity of phishing URL detention augmented with Threat Intelligence. The MLSELM ensemble features XGboost and Random Forest as complementary to each other and to obtain more exclusive decision boundary. It is this combined design that makes it possible to detect zero-day phishing URLs through adaptive mechanisms and also enhances prediction confidence. The functionality of the Chrome Extension to use PhishTank and OpenPhish feeds also confirms suspicious inputs to ensure that the accuracy of the system remains high during actual use.

The performance of each of the classifiers were measured using accuracy, precision, recall and F1-score, where phishing URLs were treated as the positive and authentic URLs as the negative used. These metrics indicate the ability of the model to distinguish safe and malicious links. Lower values give cause to misclassification and higher values give stronger indications of discrimination and reliability. It was seen that the inclusion of TI verification gave a higher number of correct decisions and a reduced number of false negative, indicating that the hybrid ML-TI architecture had a better robustness when it came to practice phishing detection.

The performance metrics were defined as follows:

1) True Positive (TP) Phishing URLs are classified as phishing correctly.
2) True Negative (TN) Legitimate URLs are classified as legitimate correctly.
3) False Positive (FP) Legitimate URLs have been classified as phishing unreasonably.
4) False Negative (FN) Phishing URLs classified casually.

- Each and every computation done on the metric is defined as: Accuracy: Accuracy is defined as the ratio of the number of true positives (those who are correctly identified as positive) to the total number of cases examined.
  ((TP +TN)/ (P +N)) × 100 ------ > 1
- Precision: The ratio of TP to the number of Ps (positives) identified.
  (TP/ (TP + FP)) × 100 ----------> 2
- Recall: The ratio of true positives to all data's positive cases.
  (TP/ (TP + FN)) × 100 ----------> 3
- F1 Score: This score combines both precision and recall in one measure:
  (2 × Precision × Recall)/ (Precision + Recall) × 100. ----------> 4

The models MLSELM, XGboost, and Random Forest were trained on balanced datasets from Kaggle and UCI ML Repository using TF-IDF feature extraction for uniform representation. All models followed the same processing pipeline for fair evaluation. As shown in Table 1, the hybrid MLSELM with Threat Intelligence achieved the best results with an accuracy of 98.9%, outperforming the individual classifiers in precision and recall.

Although the ensemble showed accuracy as competitors of both XGboost and Random Forest, they are more successful in identifying the zero-day phishing URLs. The hybrid model minimized false negative and offered greater flexibility by incorporating real-time feeds of Threat Intelligence. It is all well, but the suggested system will provide its users with enhanced safety and timely notifications and assistance, as it is the reliable, timely Real-time phishing detection via Chrome Extension.

MODEL	ACCURACY (%)	PRECISION (%)	RECALL (%)	F1 SCORE (%)
MLSELM	98.9	98.7	98.9	98.8
XG Boost	97.4	96.9	97.1	97.0
Random Forest	96.8	96.2	96.4	96.3

Table 1 : Model Performance

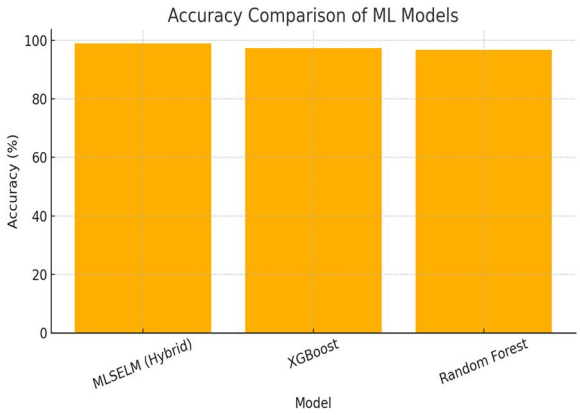


Figure 4 – Model Accuracy plot

The graph shows that the MLSELM hybrid model achieves the highest accuracy among all three classifiers. XG Boost perform slightly lower followed closely by random forest. The result confirms that combining ML model with threat intelligent significantly improves overall detection accuracy.

V. INFERENCES

The Hybrid MLSELM model is a stacked ensemble model that merges both the benefits of random forest and xgboost to produce better phishing detection results. It has been enhanced with Threat Intelligence (TI) and uses the Live Phishing URL detection feeds of PhishTank and OpenPhish to identify zero-day phishing URLs. This fusion is more flexible and reliable in decision-making than the traditional one-layered models.

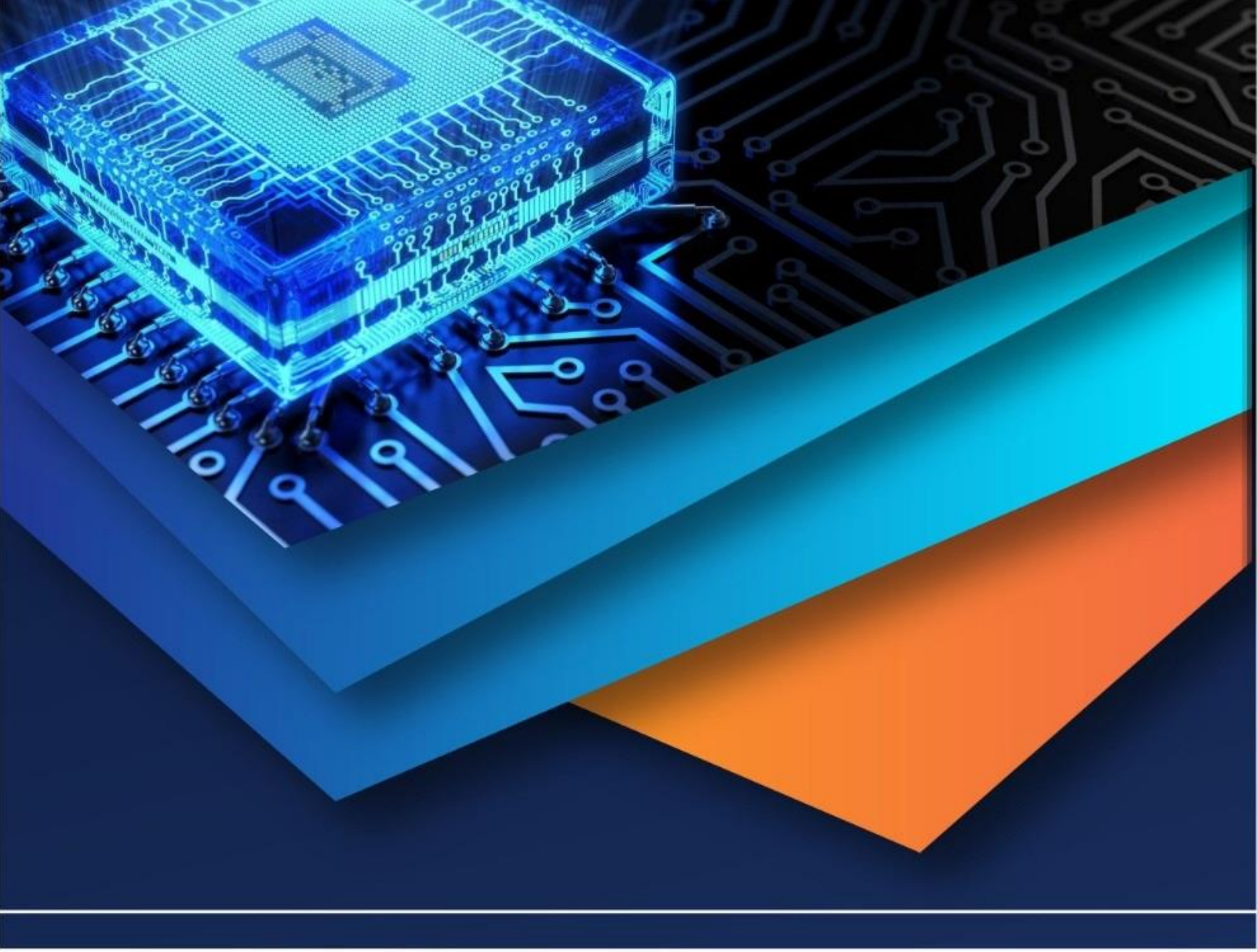
TF-IDF feature extraction method adds more strength to the system whereby the URLs are converted into meaningful numerical patterns. These features and real-time TI data enable the model to reduce false results but at the same time identify the concealed phishing features. Everything said and done, hybrid ML -TI architecture will ensure faster, more accurate, and dependable phishing detections in order to protect against real-life browsing.

VI. CONCLUSION

This paper proposes a Hybrid MLSELM Threat intelligent system designed to detect phishing attacks with high accuracy by integrated ensemble machine learning with live threat verification feeds to deliver strong precision and resilience against zero-day phishing attempts. implemented as a chrome extension it provides users with real time reliable protection during wed browsing. The result demonstrate that the hybrid intelligence approach significance improves both detection precision and overall cybersecurity effectiveness

REFERENCES

[1] M. Almousa and M. Anwar, "A URL-based social semantic attacks detection with character-aware language model," *IEEE Access*, vol. 11, Jan. 2023.
[2] M. Ok, I. Kara, and A. Ozaday, "Characteristics of understanding URLs and domain names features: The detection of phishing websites with machine learning methods," *IEEE Access*, vol. 10, Nov. 2022.
[3] O. K. Sahingoz, E. Buber, and E. Kugu, "DEPHIDES: Deep learning based phishing detection system," *IEEE Access*, vol. 12, Jan. 2024
[4] A. Karim, S. B. Belhaouari, M. Shahroz, K. Mustofa, and S. R. Joga, "Phishing detection system through hybrid machine learning based on URL," *IEEE Access*, vol. 11, Mar. 2023.
[5] S. Bell and P. Komisarczuk, "An analysis of phishing blacklists: Google Safe Browsing, OpenPhish, and PhishTank," in Proc. Australas. Comput. Sci. Week Multiconf. (ACSW), Melbourne, VIC, Australia. New York, NY, USA: Association for Computing Machinery, 2020, Art. no. 3.
[6] M. Aminu, O. Oyedokun, A. Akinsanya, and D. Apaleokhai, "Enhancing Cyber Threat Detection through Real-time Threat Intelligence and Adaptive Defense Mechanisms," Int. J. Comput. Appl. Technol. Res., vol. 13, no. 8, 2024.
[7] L. Tang and Q. H. Mahmoud, "A deep learning-based framework for phishing website detection," in Proc. IEEE, 2023.
[8] R. Zieni, L. Massari, and M. C. Calzarossa, "Phishing or not phishing? A survey on the detection of phishing websites," *IEEE Access*, vol. 11, ,2023.
[9] C. Ma et al., "Phishing website detection based on deep learning technique," IEEE Access, vol. 8, 2020
[10] "Staying Ahead of Phishers: A Review of Recent Advances and Phishing Detection," Appl. Intell., Springer, 2024.
[11] P. Maneriker, J. W. Stokes, E. G. Lazo, D. Carutasu, F. Tajaddodianfar, and A. Gururajan, "URLTran: Improving phishing URL detection using transformers," IEEE, Jun. 2021.
[12] W. Guo, Q. Wang, H. Yue, H. Sun, and R. Q. Hu, "Efficient phishing URL detection using graph-based machine learning and loopy belief propagation," IEEE, Jan. 2025.
[13] B. T. Mummadi and N. Puligundla, "Detection of phishing websites using supervised learning," Int. J. Intell. Syst. Appl. Eng., 2022.
[14] "Unveiling Suspicious Phishing Attacks: Enhancing Detection with an Optimal Feature Vectorization Algorithm," Frontiers in Computer Science, 2024.
[15] M. A. Daniel, S.-C. Chong, L.-Y. Chong, and K.-K. Wee, "Optimising phishing detection: A comparative analysis of machine learning methods with feature selection," J. Informatics Web Eng., Feb. 2025.
[16] Wenhao Li, S. Manickam, Y.-W. Chong, W. Leng, and P. Nanda, "A state-of-the-art review on phishing website detection techniques," IEEE Access, vol. 12, Dec. 2024.
[17] Machine Learning Techniques for Phishing Detection: Strengthening Threat Intelligence Sharing Mechanisms," Security Journal, 2025.
[18] D. J. Dsouza, A. P. Rodrigues, and R. Fernandes, "Multi- modal comparative analysis on execution of phishing detection using artificial intelligence," IEEE Access, vol. 12, Nov. 2024.
[19] A. Author1, B. Author2, and C. Author3, "BGL- PhishNet: Phishing website detection using hybrid model— BERT, GNN, and LightGBM," IEEE Access, vol. 11, Dec. 2023
[20] B. Lim, R. Huerta, A. Sotelo, et al., "EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability," arXiv preprint arXiv:2503.20796, 2025
[21] A. Jain and V. Gupta, "A hybrid CNN-XGBoost model for phishing URL classification," IEEE Access, vol. 12, 2024.
[22] H. AlEroud and M. Karabatis, "Integrating threat intelligence with machine learning for cyberattack detection," Computers & Security, Elsevier, 2023.
[23] Z. Zhou, L. Han, and P. Kumar, "Real-time phishing detection using transformer-based feature encoding," Expert Systems With Applications, vol. 236, 2024.
[24] R. Shanthamallu et al., "Ensemble learning for cybersecurity threat detection: A survey," ACM Computing Surveys, 2023
[25] A. Singh and M. Sharma, "A lightweight phishing detection model for browser extensions using ML and TI feeds," Springer: Neural Computing & Applications, 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY